Estimation Theory of Semi-Supervised Learning : From the Geometry of Influence Function to Construction of Estimating Equation (New developments of geometry of statistical manifolds)

AUTHOR(S):

川喜田, 雅則

CITATION:

川喜田, 雅則. Estimation Theory of Semi-Supervised Learning : From the Geometry of Influence Function to Construction of Estimating Equation (New developments of geometry of statistical manifolds). 数理解析研究所講究録 2014, 1916: 123-142: KJ00009510319.

ISSUE DATE:

2014-09

URL:

http://hdl.handle.net/2433/223323

数理解析研究所講究録
第1916巻 2014年 123-142

Estimation Theory of Semi-Supervised Learning -From the Geometry of Influence Function to Construction of Estimating Equation-

Masanori Kawakita
Graduate School of Information Science and Electrical Engineering
Kyushu University

May 7, 2014

Abstract

In the past, we proposed a safe semi-supervised learning based on weighted likelihood. We also showed that our method improves the supervised learning asymptotically in view of information geometry. In this study, we discuss its converse. If we are given a geometrical object (influence function), can we recover its associated estimator? As a result, we derive a set of all influence functions of regular and asymptotically linear estimators in semi-supervised setting. In this analysis, we do not assume that the model of $p(y|x)$ is correctly specified. Next, given an influence function, we reconstruct an estimating function such that the resultant estimator has the given influence function.

1 Semi-Supervised Learning

We review the setting of semi-supervised learning. Suppose that we have a data set:

$$\begin{aligned}(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) &\sim p(x,y), \text{ i.i.d.}\\ x'_1,x'_2,\cdots,x'_{n'} &\sim p(x')=\int p(x',y)dy. \text{ i.i.d.}\end{aligned}$$

Our interest is to estimate the conditional distribution $p(y|x)$. For this purpose, we prepare two models:

$$\begin{aligned}\mathscr{M}_x &:= \{g(x;\eta)\,|\,\eta\in\Re^{d'}\}\\ \mathscr{M} &:= \{p(y|x;\alpha)\,|\,\alpha\in\Re^{d}\}.\end{aligned}$$

The goal of semi-supervised learning is to estimate α such that $p(y|x;\alpha)$ is as close to $p(y|x)$ as possible in a certain distance measure.

2 Notations

We introduce several notations used in this paper. Let u_α, u_η be some estimating functions of parameter α and η and s_α and s_η be score functions with the models $\mathscr{M}_x$ and $\mathscr{M}$. We define

the following symbols:

$$s_\eta := s_\eta(x;\eta_0), s'_\eta := s_\eta(x';\eta_0), u_\eta := s_\eta(x;\eta_0), u'_\eta := s_\eta(x';\eta_0), u_\alpha := u_\alpha(x';\alpha_0),$$
$$r := n/n', J_{\tilde{\alpha}} := -E_{p(x,y)}\left[\frac{\partial u_\alpha(X,Y;\alpha_0)}{\partial \alpha^T}\right], J_{\tilde{\eta}} := -E_{p(x,y)}\left[\frac{\partial u_\eta(X,Y;\eta_0)}{\partial \eta^T}\right], \theta = (\alpha^T, \eta^T)^T,$$
$$G_{\tilde{\alpha}\tilde{\alpha}} := E_{p(x,y)}\left[u_\alpha(X,Y;\alpha_0)u_\alpha(X,Y;\alpha_0)^T\right], G_{\tilde{\eta}\tilde{\eta}} := E_{p(x,y)}\left[u_\eta(X,Y;\eta_0)u_\eta(X,Y;\eta_0)^T\right],$$
$$G_{\tilde{\alpha}\eta} := E[u_\alpha(X,Y;\alpha_0)s_\eta(X;\eta_0)^T], G_{\tilde{\alpha}\alpha} := E[u_\alpha(X,Y;\alpha_0)s_\alpha(Y|X;\alpha_0)^T].$$

As for $G_{\cdot\cdot}$, we listed only two examples above. The rule is to use tilde to express a general estimating function. Without tilde, it indicates a score function. We also use the following acronyms:

IF Influence Function

SIF Semi-supervised Influence Function

IIF Interest Influence Function

NIF Nuisance Influence Function.

3 Background

Let us consider an arbitrary supervised estimator defined as

$$\text{supervised estimator} \qquad \hat{\alpha} = \underset{\alpha\in\Re^d}{\text{argsolve}}\left\{\sum_{i=1}^n u_\alpha(x_i,y_i;\alpha) = 0\right\},$$
$$\text{supervised influence function} \qquad \sqrt{n}(\hat{\alpha}-\alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n J_{\tilde{\alpha}}^{-1}u_\alpha(x_i,y_i;\alpha_0) + o_p(1).$$

where $u_\alpha(x,y;\alpha)$ is an estimating function (i.e., $E_{g(x;\eta)p(y|x;\alpha)}[u(X,Y;\alpha)] = 0$ for any θ). Further, $\alpha_0 := \text{argsolve}_\alpha\{E[u_\alpha(X,Y;\alpha)] = 0\}$. Kawakita and Takeuchi (2014) showed that DRESS I improves the supervised estimator if $g(x;\eta)$ is a correct model and $p(y|x;\alpha)$ is a wrong model. DRESS I is defined as the solution of the following estimating equations:

$$\begin{aligned}
\text{DRESS I} \quad \tilde{\alpha} &= \underset{\alpha\in\Re^d}{\text{argsolve}}\left\{\sum_{i=1}^n u_\alpha(x_i,y_i;\alpha)w(x_i;\hat{\eta},\hat{\eta}') = 0\right\},\\
\hat{\eta} &= \underset{\eta\in\Re^{d'}}{\text{argsolve}}\left\{\sum_{i=1}^n s_\eta(x_i;\eta) = 0\right\},\\
\hat{\eta}' &= \underset{\eta'\in\Re^{d'}}{\text{argsolve}}\left\{\sum_{j=1}^{n'} s_\eta(x'_j;\eta') = 0\right\}.
\end{aligned} \tag{1}$$

where $w(x;\eta,\eta') = g(x;\eta')/g(x;\eta)$. Its influence function was also elucidated as

$$\begin{aligned}
\text{IIF} \quad \sqrt{n}(\hat{\alpha}-\alpha_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(J_{\tilde{\alpha}}^{-1}u_\alpha(x_i,y_i;\alpha_0) - J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s_\eta(x_i;\eta_0)\right) \\
&\quad + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'}\sqrt{r}J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s(x'_j;\eta_0) + o_p(1), \\
\text{NIF} \quad \sqrt{n}(\hat{\eta}-\eta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}G_{\eta\eta}^{-1}s_\eta(x_i;\eta_0) + o_p(1), \\
\text{NIF} \quad \sqrt{n}(\hat{\eta}'-\eta_0) &= \sqrt{r}\frac{1}{\sqrt{n}}\sum_{j=1}^{n'}G_{\eta\eta}^{-1}s_\eta(x'_j;\eta_0) + o_p(1).
\end{aligned} \tag{2}$$

Kawakita and Takeuchi (2014) showed that IIF can be interpreted as

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = J_{\tilde{\alpha}}^{-1}u_\alpha - (1+r)\Pi[J_{\tilde{\alpha}}^{-1}u_\alpha | s_\eta - \sqrt{r}s'_\eta].$$

This can be illustrated geometrically as in Fig. 1. In this figure, DRESS I corresponds to the

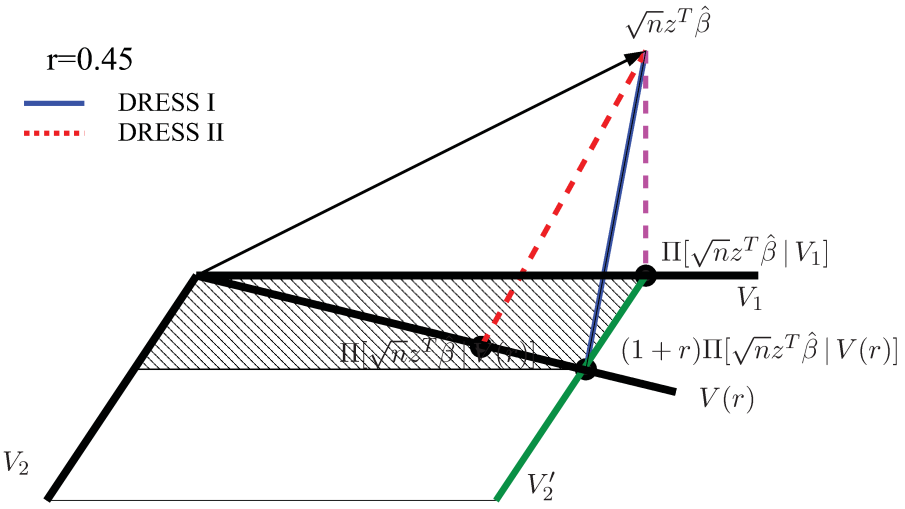

Figure 1: Geometry of estimators.

blue line. It is easy to see that the blue line is always shorter than the black line (supervised estimator) unless $r \leq 1$. As in Kawakita and Takeuchi (2014), an existence of better estimator is suggested from this figure. That is a red line. Its IF is given by

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = J_{\tilde{\alpha}}^{-1}u_\alpha - \Pi[J_{\tilde{\alpha}}^{-1}u_\alpha | s_\eta - \sqrt{r}s'_\eta].$$

However, we are not sure whether such an estimator exists or how to obtain it? In (Kawakita and Takeuchi, 2014), such an estimator was heuristically found. That is DRESS II estimator

defined as

$$\begin{aligned}
\text{DRESS II} \quad \tilde{\alpha} &= \underset{\alpha\in\Re^d}{\text{argsolve}} \left\{ \sum_{i=1}^{n} u_\alpha(x_i, y_i; \alpha) w(x_i; \hat{\eta}, \hat{\eta}') = 0 \right\}, \\
\hat{\eta} &= \underset{\eta\in\Re^{d'}}{\text{argsolve}} \left\{ \sum_{i=1}^{n} s_\eta(x_i; \eta) = 0 \right\}, \\
\hat{\eta}' &= \underset{\eta'\in\Re^{d'}}{\text{argsolve}} \left\{ \sum_{i=1}^{n} s_\eta(x_i; \eta') + \sum_{j=1}^{n'} s_\eta(x'_j; \eta') = 0 \right\}. \qquad (3)
\end{aligned}$$

Its influence function is given by

$$\begin{aligned}
\text{IIF} \quad \sqrt{n}(\hat{\alpha} - \alpha_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J_{\tilde{\alpha}}^{-1} u_\alpha(x_i, y_i; \alpha_0) - J_{\tilde{\alpha}}^{-1} G_{\tilde{\alpha}\eta} G_{\eta\eta}^{-1} s_\eta(x_i; \eta_0) \qquad (4) \\
&\quad + J_{\tilde{\alpha}}^{-1} G_{\tilde{\alpha}\eta} G_{\eta\eta}^{-1} s(x'_j; \eta_0) + o_p(1), \\
\text{NIF} \quad \sqrt{n}(\hat{\eta} - \eta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} G_{\eta\eta}^{-1} s_\eta(x_i; \eta_0) + o_p(1), \\
\text{NIF} \quad \sqrt{n}(\hat{\eta}' - \eta_0) &= \frac{r}{r+1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} G_{\eta\eta}^{-1} s_\eta(x_i; \eta_0) + \frac{\sqrt{r}}{r+1} \frac{1}{\sqrt{n'}} \sum_{j=1}^{n'} G_{\eta\eta}^{-1} s_\eta(x'_j; \eta_0) + o_p(1). \qquad (5)
\end{aligned}$$

Our goal is to establish the method to do these. More concretely, we aim at giving answers to the following questions:

- For each IF, is there an estimator corresponding to it?
- If it exists, how can we obtain it?

3.1 Problem formulation

Suppose that

$$\begin{aligned}
(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) &\sim p(x)p(y|x), \text{ i.i.d.} \\
x'_1, x'_2, \cdots, x'_{n'} &\sim p(x'), \text{ i.i.d..}
\end{aligned}$$

Further, suppose that we are given estimating functions

estimating function for α $\quad u_\alpha(x, y; \alpha)$

estimating function for η $\quad u_\eta(x; \eta)$ $\quad$ (this is not necessary. just a virtual concept.)

Define two target parameter values as the the function of $p(x, y)$:

$$\begin{aligned}
\alpha(p) &:= \underset{\alpha\in\Re^d}{\text{argsolve}} \left\{ E_{p(x,y)}[u_\alpha(X, Y; \alpha)] = \mathbf{0} \right\} \qquad (6) \\
\eta(p) &= \underset{\eta\in\Re^{d'}}{\text{argsolve}} \left\{ E_{p(x)}[u_\eta(X; \eta)] = \mathbf{0} \right\}. \qquad (7)
\end{aligned}$$

Note that $\alpha(p)$ clearly depends on u_α. Our goal is

- to elucidate the set of all IIF of regular and asymptotically linear estimators of $\hat{\alpha}$,
- for any fixed IIF, to derive the set of estimating equations yielding the estimator which has the given IIF.

3.2 Class of all influence functions

Again, we focus on the regular estimators.

Definition 1 (local data generating process). *Let $\{p_n(y|x)\}$ and $\{p_n(x)\}$ be sequences of conditional and marginal probability density functions such that*

$$\begin{aligned}(x_{n1}, y_{n1}), (x_{n2}, y_{n2}), \cdots, (x_{nn}, y_{nn}) &\sim p_n(x, y), \ i.i.d. \\ x'_{n1}, x'_{n2}, \cdots, x'_{nn'} &\sim p'_n(x') = \int p_n(x', y)dy, \ i.i.d.\end{aligned}$$

where $p_n(x, y) := p_n(x)p_n(y|x)$. Define

$$\begin{aligned}\delta_n(x) &:= \sqrt{n}\left(\frac{p_n(x) - p(x)}{p(x)}\right) \\ \gamma_n(x, y) &:= \sqrt{n}\left(\frac{p_n(y|x) - p(y|x)}{p(y|x)}\right).\end{aligned}$$

We say $(p_n(x), p_n(y|x))$ is a local data generating process if $\delta_n(x)$ and $\gamma_n(x, y)$ converge to some functions.

We write $\alpha(p_n(x, y))$, $\alpha(p(x, y))$, $\eta(p_n(x))$ and $\eta(p(x))$ as $\alpha_n, \alpha_0, \eta_n, \eta_0$ respectively. For any LDGP, the following lemma holds.

Lemma 2. *For any LDGP, $\sqrt{n}(\alpha_n - \alpha_0)$ and $\sqrt{n}(\eta_n - \eta_0)$ converge to some constant vectors.*

Proof. Assume that $p_n(x)$ and $p_n(x, y)$ are LDGP. Then, we have

$$\begin{aligned}p_n(x) &= p(x)\left(1 + \frac{1}{\sqrt{n}}\delta_n(x)\right), \\ p_n(y|x) &= p(y|x)\left(1 + \frac{1}{\sqrt{n}}\gamma_n(x, y)\right), \\ p_n(x, y) &:= p_n(x)p_n(y|x) = p(x, y)\left(1 + \frac{1}{\sqrt{n}}\delta_n(x) + \frac{1}{\sqrt{n}}\gamma_n(x, y)\right) + O(1/n).\end{aligned}$$

By their definitions, it holds that

$$\begin{aligned}E_{p_n}[u_\alpha(X, Y; \alpha_n)] &= \mathbf{0}, \\ E_{p_n}[u_\eta(X; \eta_n)] &= \mathbf{0}.\end{aligned}$$

By Taylor-expansion, we have

$$\begin{aligned}E_{p_n}\left[u_\alpha(X, Y; \alpha_0) + \frac{\partial u_\alpha}{\partial \alpha^T}(\alpha_n - \alpha_0) + o(1/\sqrt{n})\right] &= \mathbf{0}, \\ E_{p_n}\left[u_\eta(X; \eta_0) + \frac{\partial u_\eta}{\partial \eta^T}(\eta_n - \eta_0) + o(1/\sqrt{n})\right] &= \mathbf{0}.\end{aligned}$$

Multiplying $\sqrt{n}$,

$$\begin{aligned}E_{p_n}\left[\sqrt{n}u_\alpha(X, Y; \alpha_0) + \frac{\partial u_\alpha}{\partial \alpha^T}\sqrt{n}(\alpha_n - \alpha_0) + o(1)\right] &= \mathbf{0}, \\ E_{p_n}\left[\sqrt{n}u_\eta(X; \eta_0) + \frac{\partial u_\eta}{\partial \eta^T}\sqrt{n}(\eta_n - \eta_0) + o(1)\right] &= \mathbf{0}.\end{aligned} \tag{8}$$

Under the assumption that u_α and u_η are of C^1 class, we have

$$E_{p_n}\left[\frac{\partial u_\alpha}{\partial \alpha^T}\right] \to E_p\left[\frac{\partial u_\alpha}{\partial \alpha^T}\right], \; E_{p_n}\left[\frac{\partial u_\eta}{\partial \eta^T}\right] \to E_p\left[\frac{\partial u_\eta}{\partial \eta^T}\right].$$

The first term is evaluated as

$$\begin{aligned}\sqrt{n}E_{p_n}\left[u_\eta(X;\eta_0)\right] &= \sqrt{n}\int p(x)\left(1+\frac{1}{\sqrt{n}}\delta_n(x)\right)u_\eta(x;\eta_0)dxdy \\ &= \int p(x)\delta_n(x)u_\eta(x;\eta_0)dxdy \\ &= \int p(x)\Pi[\delta_n(x)|u_\eta]u_\eta(x;\eta_0)dxdy\end{aligned}$$

Define $b'_n := G^{-1}_{\tilde{\eta}\tilde{\eta}}E_{p(x)}[\delta_n(x)u_\eta(X;\eta_0)]$. Then, $\Pi[\delta_n(x)|u_\eta] = u_\eta^T b'_n$. Using this,

$$\sqrt{n}E_{p_n}\left[u_\eta(X;\eta_0)\right] = G_{\tilde{\eta}\tilde{\eta}}b'_n.$$

As a result, we have

$$G_{\tilde{\eta}\tilde{\eta}}b'_n - J_{\tilde{\eta}}\sqrt{n}(\eta_n - \eta_0) + o(1) = \mathbf{0}.$$

Hence, it holds that

$$\sqrt{n}(\eta_n - \eta_0) = J^{-1}_{\tilde{\eta}}G_{\tilde{\eta}\tilde{\eta}}b'_n + o(1).$$

Because $\delta_n(x)$ converges to a some function, $\{b'_n\}$ also converges to a some constant vector b' so that

$$\sqrt{n}(\eta_n - \eta_0) \to J^{-1}_{\tilde{\eta}}G_{\tilde{\eta}\tilde{\eta}}b' + o(1).$$

Similarly, the other first term is evaluated as

$$\begin{aligned}\sqrt{n}E_{p_n}\left[u_\alpha(X,Y;\alpha_0)\right] &= \sqrt{n}\int p(x,y)\left(1+\frac{1}{\sqrt{n}}\delta_n(x)+\frac{1}{\sqrt{n}}\gamma_n(x,y)+O(1/n)\right)\cdot \\ &\quad u_\alpha(x,y;\alpha_0)dxdy \\ &= \int p(x,y)\left(\delta_n(x)+\gamma_n(x,y)\right)u_\alpha(x,y;\alpha_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right). \quad (9)\end{aligned}$$

The orthogonal projection of $\delta_n(x)+\gamma_n(x,y)$ onto u_α is decomposed as

$$\Pi[\delta_n(x)+\gamma_n(x,y)|u_\alpha] = \Pi[\delta_n(x)|u_\alpha] + \Pi[\gamma_n(x,y)|u_\alpha].$$

Let $b_n := G_{\tilde{\alpha}\tilde{\alpha}}E_{p(x,y)}[\delta(x,y)u_\alpha(X,Y;\alpha_0)]$ so that $\Pi[\gamma_n(x,y)|u_\alpha] = b_n^T u_\alpha$. On the other hand, $\delta_n(x)$ is decomposed into $\delta_n(x) = u_\eta^T b'_n + (\delta_n(x) - u_\eta^T b'_n)$. We write $(\delta_n(x) - u_\eta^T b'_n)$ as $h_n(x)$ for short. We can further decompose $\delta_n(x)$ as

$$\begin{aligned}\delta_n(x) &= \Pi[u_\eta^T b'_n + h_n|u_\alpha] + (\delta_n(x) - \Pi[\delta_n(x)|u_\alpha]) \\ &= (G^{-1}_{\tilde{\alpha}\tilde{\alpha}}G_{\tilde{\alpha}\tilde{\eta}}b'_n)^T u_\alpha + \Pi[h_n|u_\alpha] + (\delta_n(x) - \Pi[\delta_n(x)|u_\alpha])\end{aligned}$$

Let $c_n := G^{-1}_{\tilde{\alpha}\tilde{\alpha}}E_{p(x)}[h_n u_\alpha]$ so that $\Pi[h_n|u_\alpha] = c_n^T u_\alpha$. As a result,

$$\delta_n(x) = (G^{-1}_{\tilde{\alpha}\tilde{\alpha}}G_{\tilde{\alpha}\tilde{\eta}}b'_n + c_n)^T u_\alpha + (\delta_n(x) - \Pi[\delta_n(x)|u_\alpha]).$$

Note that c_n also converges to a some constant c because h_n converges to a some function. Similarly, let $b_n := G_{\tilde{\alpha}\tilde{\alpha}}^{-1} E_{p(x,y)}[\gamma_n(x,y)u_\alpha]$ so $\Pi[\gamma_n(x,y)|u_\alpha] = b_n^T u_\alpha$. Because of the same reason again, b_n converges to a some constant b. Summarizing these, we have

$$\Pi[\delta_n(x) + \gamma_n(x,y)|u_\alpha] = (G_{\tilde{\alpha}\tilde{\alpha}}^{-1} G_{\tilde{\alpha}\tilde{\eta}} b_n' + c_n + b_n)^T u_\alpha.$$

Substituting this into Eq. (9), we have

$$\begin{aligned}
\sqrt{n}E_{p_n}\left[u_\alpha(X,Y;\alpha_0)\right] &= \int p(x,y)\left(\delta_n(x) + \gamma_n(x,y)\right) u_\alpha(x,y;\alpha_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right) \\
&= \int p(x,y)\Pi[\left(\delta_n(x) + \gamma_n(x,y)\right)|u_\alpha]u_\alpha(x,y;\alpha_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right) \\
&= G_{\tilde{\alpha}\tilde{\alpha}}(G_{\tilde{\alpha}\tilde{\alpha}}^{-1} G_{\tilde{\eta}\tilde{\alpha}} b_n' + c_n + b_n) + O\left(\frac{1}{\sqrt{n}}\right) \\
&= (G_{\tilde{\eta}\tilde{\alpha}} b_n' + G_{\tilde{\alpha}\tilde{\alpha}}(c_n + b_n)) + O\left(\frac{1}{\sqrt{n}}\right) \\
&\to (G_{\tilde{\eta}\tilde{\alpha}} b' + G_{\tilde{\alpha}\tilde{\alpha}}(c + b)) + O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Hence, from Eq. (8),

$$\sqrt{n}(\alpha_n - \alpha_0) \to J_{\tilde{\alpha}}^{-1}(G_{\tilde{\alpha}\tilde{\eta}} b' + G_{\tilde{\alpha}\tilde{\alpha}}(c + b)).$$

□

Lemma 3. *For any LDGP $(p_n(x), p_n(y|x))$, $\delta_n(x)$ and $\gamma_n(x,y)$ satisfy the following properties:*

1. *Both $\delta_n(x)$ and $\gamma_n(x,y)$ have zero mean.*
2. *The function $\delta_n(x)$ is perpendicular to any function $a(x,y)$ such that $E_{p(y|x)}[a(X,Y)]$=0.*
3. *The function $\gamma_n(x,y)$ is perpendicular to any function $a(x)$.*

Proof. Taking expectation, we have

$$\begin{aligned}
\int p(x)\delta_n(x)dx &= \int p(x)\sqrt{n}\left(\frac{p_n(x) - p(x)}{p(x)}\right)dx = \sqrt{n}\int (p_n(x) - p(x))dx = 0, \\
\int p(x,y)\gamma_n(x,y)dxdy &= \int p(x,y)\sqrt{n}\left(\frac{p_n(y|x) - p(y|x)}{p(y|x)}\right)dx \\
&= \sqrt{n}\int p(x)(p_n(y|x) - p(y|x))dxdy = 0.
\end{aligned}$$

For any function $a(x,y)$,

$$E_p[\delta_n(X)a(X,Y)] = \int p(x,y)\delta_n(x)a(x,y)dxdy = \int p(x)\delta_n(x)\int p(y|x)a(x,y)dydx.$$

Therefore, if $\int p(y|x)a(x,y)dy = 0$, then δ_n is perpendicular to a. For any function $a(x)$,

$$E_p[\gamma_n(X,Y)a(X)] = \int p(x,y)a(x)\gamma_n(x,y) = \int p(x)a(x)\int p(y|x)\gamma_n(x,y)dydx = 0.$$

□

From this lemma, we immediately know that δ_n is perpendicular to γ_n.

Lemma 4. *Consider an LDGP $(p_n(x), p_n(y|x))$. Define two probability density sequences:*

$$\begin{aligned} \textit{truth} \quad & p_{0n}(v_n) := \Pi_{i=1}^{n} p(x_i, y_i) \Pi_{j=1}^{n'} p(x'_j), \\ \textit{LDGP} \quad & p_{1n}(v_n) := \Pi_{i=1}^{n} p_n(x_i, y_i) \Pi_{j=1}^{n'} p_n(x'_j). \end{aligned}$$

If and only if $E_{p(x)}[\frac{p_n(x)}{p(x)}\frac{p_n(x)}{p(x)}]$ and $E_{p(y|x)}[\frac{p_n(y|x)}{p(y|x)}\frac{p_n(y|x)}{p(y|x)}]$ are finite, then, $E_{p(x)}[\delta_n(X)^2]$ and $E_{p(x,y)}[\gamma_n(X,Y)^2]$ are finite for any n.

Proof. By definition, we have

$$\begin{aligned} E_{p(x,y)}[\delta_n(X)^2] &= \int p(x) \left(\sqrt{n} \frac{(p_n(x) - p(x))}{p(x)} \right)^2 dx \\ &= n \int \left(\frac{(p_n(x)^2 - 2p_n(x)p(x) + p(x)^2)}{p(x)} \right) dx \\ &= n \int \left(p(x) \left(\frac{p_n(x)^2}{p(x)^2} \right) - 2p_n(x) + p(x)) \right) dx \\ &= n \left(\int p(x) \left(\frac{p_n(x)^2}{p(x)^2} \right) dx - 1 \right). \end{aligned}$$

Therefore, the finiteness of the squared expectation norm of p_n/p is equivalent to the finiteness of $E[\delta_n^2]$. Similarly, we have

$$\begin{aligned} E_{p(x,y)}[\gamma_n(X,Y)^2] &= \int p(x,y) \left(\sqrt{n} \frac{(p_n(y|x) - p(y|x))}{p(y|x)} \right)^2 dxdy \\ &= n \int \left(\frac{(p_n(y|x)^2 - 2p_n(y|x)p(y|x) + p(x)^2)}{p(y|x)} \right) dxdy \\ &= n \int \cdot \left(p(y|x) \left(\frac{p_n(y|x)^2}{p(y|x)^2} \right) - 2p_n(x) + p(y|x)) \right) dxdy \\ &= n \left(\int p(x) \left(\frac{p_n(x)^2}{p(x)^2} \right) dx - 1 \right). \end{aligned}$$

□

Lemma 5. *Consider an LDGP $(p_n(x), p_n(y|x))$. Define two probability density sequences:*

$$\begin{aligned} \textit{truth} \quad & p_{0n}(v_n) := \Pi_{i=1}^{n} p(x_i, y_i) \Pi_{j=1}^{n'} p(x'_j), \\ \textit{LDGP} \quad & p_{1n}(v_n) := \Pi_{i=1}^{n} p_n(x_i, y_i) \Pi_{j=1}^{n'} p_n(x'_j). \end{aligned}$$

If $E_{p(x)}[\frac{p_n(x)}{p(x)}\frac{p_n(x)}{p(x)}]$ and $E_{p(y|x)}[\frac{p_n(y|x)}{p(y|x)}\frac{p_n(y|x)}{p(y|x)}]$ are finite, then, p_{1n} is contiguous to p_{0n}.

Proof. By definition, we have

$$\begin{aligned}
&\log\left(\frac{p_{1n}(v_n)}{p_{0n}(v_n)}\right)\\
=\ &\log\left(\frac{\Pi_{i=1}^n p_n(x_i,y_i)\Pi_{j=1}^{n'} p_n(x'_j)}{\Pi_{i=1}^n p(x_i,y_i)\Pi_{j=1}^{n'} p(x'_j)}\right)\\
=\ &\log\left(\Pi_{i=1}^n p(x_i,y_i)\left(1+\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)\right.\\
&\left.\cdot\Pi_{j=1}^{n'} p(x'_j)\left(1+\frac{1}{\sqrt{n}}\delta_n(x'_j)\right)\right)-\log\left(\Pi_{i=1}^n p(x_i,y_i)\Pi_{j=1}^{n'} p(x'_j)\right)\\
=\ &\log\left(\Pi_{i=1}^n\left(1+\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)\Pi_{j=1}^{n'}\left(1+\frac{1}{\sqrt{n}}\delta_n(x'_j)\right)\right)\\
=\ &\sum_{i=1}^n\log\left(1+\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)+\sum_{j=1}^{n'}\log\left(1+\frac{1}{\sqrt{n}}\delta_n(x'_j)\right).
\end{aligned}$$

Note that by Taylor-expansion around $t=0$, there exists $c\in[0,t]$ such that

$$\log(1+t)=t-\frac{1}{2}\frac{1}{1+c}t^2.$$

Let $t=\frac{1}{\sqrt{n}}\delta_n+\frac{1}{\sqrt{n}}\gamma_n+\frac{1}{n}\delta_n\gamma_n$ so that $t\to 0$. Then, there exists a real sequence c_n such that $c_n\to 0$ and the first term is evaluated as

$$\begin{aligned}
\text{1st term}\ &=\ \sum_{i=1}^n\log\left(1+\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)\\
&=\ \sum_{i=1}^n\left(\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)\\
&\quad-\sum_{i=1}^n\frac{1}{2}\frac{1}{1+c_n}\left(\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)^2\\
&=\ \sum_{i=1}^n\left(\frac{1}{\sqrt{n}}\delta_n(x_i)+\frac{1}{\sqrt{n}}\gamma_n(x_i,y_i)+\frac{1}{n}\delta_n(x_i)\gamma_n(x_i,y_i)\right)\\
&\quad-\sum_{i=1}^n\frac{1}{2}\frac{1}{1+c_n}\left(\frac{1}{n}\delta_n(x_i)^2+\frac{1}{n}\gamma_n(x_i,y_i)^2+\frac{2}{n}\delta_n(x_i)\gamma_n(x_i,y_i)+o(1/n)\right)\\
&=\ \frac{1}{\sqrt{n}}\sum_{i=1}^n(\delta_n(x_i)+\gamma_n(x_i,y_i))+\frac{1}{n}\sum_{i=1}^n\delta_n(x_i)\gamma_n(x_i,y_i)\\
&\quad-\frac{1}{n}\sum_{i=1}^n\frac{1}{2}\frac{1}{1+c_n}\left(\delta_n(x_i)^2+\gamma_n(x_i,y_i)^2+2\delta_n(x_i)\gamma_n(x_i,y_i)\right)+o(1)\\
&\to\ \frac{1}{\sqrt{n}}\sum_{i=1}^n(\delta_n(x_i)+\gamma_n(x_i,y_i))-\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n\left(\delta_n(x_i)^2+\gamma_n(x_i,y_i)^2\right)\right)+o(1)
\end{aligned}$$

The first term converges in distribution to $N(0,E_{p(x,y)}[\delta_n^2+\gamma_n^2])$. The second term converges in probability to

$$-\frac{1}{2}(E_{p(x,y)}[\delta_n(X)^2+\gamma_n(X,Y)^2]).$$

By denoting $E_{p(x,y)}[\delta_n^2 + \gamma_n^2]$ by τ^2, $\log\left(\frac{p_{1n}(v_n)}{p_{0n}(v_n)}\right)$ converges in distribution $N(-(1/2)\tau^2, \tau^2)$. Thus, by LeCam's lemma, p_{1n} is contiguous to p_{0n}. □

Using these lemmas, we provide a theorem corresponding to Theorem 3.2 of Tsiatis (2006) under model misspecification.

Theorem 6 (semi-supervised setting). *Let $\hat{\alpha}_n$ be asymptotically linear with influence functions (ϕ_1, ϕ_2) such that $E_p[\phi_1\phi_1^T]$ and $E_p[\phi_2\phi_2^T]$ are continuous in the neighborhood of p. If $\hat{\alpha}_n$ is regular and both $E_{p(x)}[\frac{p_n(x)}{p(x)}\frac{p_n(x)}{p(x)}]$ and $E_{p(y|x)}[\frac{p_n(y|x)}{p(y|x)}\frac{p_n(y|x)}{p(y|x)}]$ are finite, then there exists a function $a(x) : \mathcal{X} \to \Re^d$ such that*

$$\phi_1(x, y) = J_{\tilde{\alpha}}^{-1} u_\alpha(x, y; \alpha_0) - \frac{1}{\sqrt{r}}\phi_2(x). \tag{10}$$

Proof. Let $(p_n(x), p_n(y|x))$ be any LDGP. By Lemma 2, $\sqrt{n}(\alpha_n - \alpha_0)$ and $\sqrt{n}(\eta_n - \eta_0)$ converge to some constant vectors. Define two probability densities

$$\begin{aligned} \text{truth} \quad & p_{0n}(v_n) := \Pi_{i=1}^n p(x_i, y_i)\Pi_{j=1}^{n'} p(x'_j), \\ \text{LDGP} \quad & p_{1n}(v_n) := \Pi_{i=1}^n p_n(x_i, y_i)\Pi_{j=1}^{n'} p_n(x'_j). \end{aligned}$$

By asymptotic linearity, it holds that

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_1(x_i, y_i) + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} \phi_2(x'_j) + o_{p_{0n}}(1).$$

By Lemma 5, p_{1n} is contiguous to p_{0n}. Therefore, $o_p(1)$ with p_{0n} is still $o_p(1)$ with p_{1n}. Further, $\phi_1(x_i, y_i; \alpha_n)$ converges to $\phi_1(x_i, y_i; \alpha_0)$ under suitable smoothness assumptions. Similarly to ϕ_2. Thus, under LDGP, we also have

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_1(x_i, y_i) + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} \phi_2(x'_j) + o_{p_{1n}}(1).$$

Using this, we have

$$\begin{aligned} \sqrt{n}(\hat{\alpha}_n - \alpha_n) &= \sqrt{n}(\hat{\alpha}_n - \alpha_0) - \sqrt{n}(\alpha_n - \alpha_0)) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_1(x_i, y_i) + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} \phi_2(x'_j) + o_{p_{1n}}(1) - \sqrt{n}(\alpha_n - \alpha_0)) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (\phi_1(x_i, y_i) - E_{p_n}[\phi_1(X, Y)]) \quad \text{(i)} \\ &\quad + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} (\phi_2(x'_j) - E_{p_n}[\phi_2(X')]) \quad \text{(ii)} \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n E_{p_n}[\phi_1(X, Y)] \quad \text{(iii)} \\ &\quad + \frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} E_{p_n}[\phi_2(X')] \quad \text{(iv)} \\ &\quad - \sqrt{n}(\alpha_n - \alpha_0)) \quad \text{(v)} \\ &\quad + o_{p_{1n}}(1). \end{aligned} \tag{11}$$

The left hand side $\sqrt{n}(\hat{\alpha}_n - \alpha(p_n))$ has a limit distribution independent of LDGP because of its regularity. Because $\phi_1 - E_{p_n}\phi_1$ has zero mean and finite variance, the term (i) converges to $N(0, E_{p_n}[\phi_1\phi_1^T])$ by CLT. By the assumption, $E_{p_n}[\phi_1\phi_1^T]$ converges to $E_p[\phi_1\phi_1^T]$. As a result, the term (i) converges to $N(0, E_p[\phi_1\phi_1^T])$ in distribution under LDGP. Similarly, the term (ii) converges to $N(0, E_p[\phi_2\phi_2^T])$. Recalling that the term (i) and the term (ii) are statistically independent, their sum is subject to $N(0, E_p[\phi_1\phi_1^T + \phi_2\phi_2^T])$. The term (iii) is calculated as

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E_{p_n}[\phi_1(X,Y)] &= \sqrt{n}\int p_n(x,y)\phi_1(x,y)dxdy \\
&= \sqrt{n}\int p(x,y)\left(1+\frac{1}{\sqrt{n}}\delta_n(x)+\frac{1}{\sqrt{n}}\gamma_n(x,y)\right)\phi_1(x,y)dxdy + O\left(\frac{1}{\sqrt{n}}\right) \\
&= \int p(x,y)\left(\delta_n(x)+\gamma_n(x,y)\right)\phi_1(x,y)dxdy + O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Similarly, the term (iv) is calculated as

$$\begin{aligned}
\frac{1}{\sqrt{n'}}\sum_{j=1}^{n'} E_{p_n}[\phi_2(X')] &= \sqrt{n'}\int p_n(x')\phi_2(x')dx' = \sqrt{n'}\int p(x')\left(1+\frac{1}{\sqrt{n}}\delta_n(x')\right)\phi_2(x')dx' \\
&= \frac{1}{\sqrt{r}}\int p(x')\delta_n(x')\phi_2(x')dx'.
\end{aligned}$$

By the process of proof of Lemma 2, the term (v) is

$$\sqrt{n}(\alpha_n - \alpha_0) = \int p(x,y)(\delta_n(x)+\gamma_n(x,y))J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right).$$

Because the right-hand side of Eq. (11) must be independent of LDGP, we have

$$\begin{aligned}
\mathbf{0} &= \int p(x,y)\left(\delta_n(x)+\gamma_n(x,y)\right)\phi_1(x,y)dxdy + \frac{1}{\sqrt{r}}\int p(x')\delta_n(x')\phi_2(x')dx' \\
&\quad - \int p(x,y)(\delta_n(x)+\gamma_n(x,y))J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right) \\
&= \int p(x,y)\delta_n(x)\left(\phi_1(x,y) - J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0) + \frac{1}{\sqrt{r}}\phi_2(x)\right)dxdy \\
&\quad + \int p(x,y)\gamma_n(x,y)\left(\phi_1(x,y) - J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0)\right)dxdy + o(1)
\end{aligned}$$

for any δ_n and γ_n sequences. The last term requires that $\phi_1 - J_{\tilde{\alpha}}^{-1}u_\alpha$ must be a function of only x. That is, there exists a function $a(x) : \mathcal{X} \to \Re^d$ such that $\phi_1(x,y) = J_{\tilde{\alpha}}^{-1}u_\alpha + a(x)$. Therefore, the second last term requires that

$$a(x) + \frac{1}{\sqrt{r}}\phi_2(x) = 0.$$

□

We can obtain the similar theorem for nuisance influence function by the same way as Theorem 6.

Theorem 7 (semi-supervised setting). *Let $\hat{\eta}_n$ be asymptotically linear with influence functions (ϕ_1, ϕ_2) such that $E_{p(x,y)}[\phi_1\phi_1^T]$ and $E_{p(x)}[\phi_2\phi_2^T]$ are continuous in the neighborhood of $p(x,y)$. If $\hat{\eta}_n$ is regular, then ϕ_1 must be a function of only x and*

$$\phi_1(x) + \frac{1}{\sqrt{r}}\phi_2(x) = J_{\tilde{\eta}}^{-1} u_\eta(x; \eta_0). \tag{12}$$

Proof. Only the difference is the term (v) in the proof of Theorem 6. From Lemma 2, Eq. (12) is replaced with

$$\begin{aligned}
\mathbf{0} &= \int p(x,y)\left(\delta_n(x) + \gamma_n(x,y)\right)\phi_1(x,y)dxdy + \frac{1}{\sqrt{r}}\int p(x')\delta_n(x')\phi_2(x')dx' \\
&\quad - \int p(x)\delta_n(x) J_{\tilde{\eta}}^{-1} u_\eta(x,y;\eta_0)dxdy + O\left(\frac{1}{\sqrt{n}}\right) \\
&= \int p(x,y)\delta_n(x)\left(\phi_1(x,y) - J_{\tilde{\eta}}^{-1}u_\eta(x;\eta_0) + \frac{1}{\sqrt{r}}\phi_2(x)\right)dxdy \\
&\quad + \int p(x,y)\gamma_n(x,y)\phi_1(x,y)dxdy + o(1).
\end{aligned}$$

The last term requires that ϕ_1 should be a function of only x. □

For a given u_η, we call the influence function satisfying Eq. (12) *a u_η-proper NIF* while any function (ν_1, ν_2) satisfying only

$$E_{p(x)}[\phi_1(X)] = E_{p(x)}[\phi_2(X)] = \mathbf{0},$$
Both ϕ_1 and ϕ_2 have the proper covariance matrix.

is just called an NIF. We show that some examples of influence function of $\hat{\alpha}$ in this setting satisfies the above conditions. Suppose that we are given $u_\alpha(x,y;\alpha)$ and $u_\eta(x;\eta) = s_\eta(x;\eta)$ by which the solution α_0 and η_0 is defined.

Example 1: supervised learning IF is given by

$$\phi_1(x,y) = J_{\tilde{\alpha}}^{-1} u_\alpha(x,y;\alpha)\,,\ \phi_2 = 0.$$

It is trivial that this IF satisfies Eq. (10).

Example 2: DRESS I IIF is given by

$$\phi_1(x,y) = J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0) - J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s_\eta(x;\eta_0)\,,\ \phi_2(x') = \sqrt{r}J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s_\eta(x';\eta_0).$$

Therefore, we have

$$\phi_1(x,y) + \frac{1}{\sqrt{r}}\phi_2(x) = J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0).$$

Example 3: DRESS II IIF is given by

$$\begin{aligned}
\phi_1(x,y) &= J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0) - \frac{1}{1+r}J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s_\eta(x';\eta_0)\,, \\
\phi_2(x') &= \frac{\sqrt{r}}{1+r}J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}G_{\eta\eta}^{-1}s_\eta(x';\eta_0).
\end{aligned}$$

Therefore, we have

$$\phi_1(x,y) + \frac{1}{\sqrt{r}}\phi_2(x) = J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0).$$

Let us see some examples where the nuisance estimator also satisfies Eq. (12).

DRESS I The first NIF is given by $(\phi_1, \phi_2) = (J_{\tilde{\eta}}^{-1} s_\eta(x; \eta_0), 0)$. Clearly, this satisfies the NIF condition. The second NIF is given by $(\phi_1, \phi_2) = (0, \sqrt{r} J_{\tilde{\eta}}^{-1} s_\eta(x'; \eta_0))$. It is also trivial to see the NIF condition is satisfied.

DRESS II The first NIF is the same as DRESS I. The second NIF is given by

$$(\phi_1, \phi_2) = \left(\frac{r}{r+1} G_{\eta\eta}^{-1} s_\eta(x; \eta_0), \frac{\sqrt{r}}{r+1} G_{\eta\eta}^{-1} s_\eta(x'; \eta_0) \right).$$

Substituting this into Eq. (12), we have

$$\left(\frac{r}{r+1} G_{\eta\eta}^{-1} s_\eta(x; \eta_0) + \frac{1}{\sqrt{r}} \frac{\sqrt{r}}{r+1} G_{\eta\eta}^{-1} s_\eta(x; \eta_0) \right) = G_{\eta\eta}^{-1} s_\eta(x; \eta_0).$$

3.3 Decomposition of influence function

In this section, we show that any IIF in semi-supervised setting can be decomposed into some IIF and NIFs. For this purpose, we provide a series of lemmas.

Lemma 8. *Let (ϕ_1, ϕ_2) be an interest influence function in semi-supervised setting. For any interest influence function $(\phi_1'(x, y), \phi_2'(x'))$, there exists a function ρ of only x such that*

$$\begin{aligned} &\rho : \mathcal{X} \to \Re^d, \\ &E_{p(x)}[\rho(X)] = 0, \\ &E_{p(x)}[\rho(X)\rho(X)^T] < \infty. \end{aligned}$$

and (ϕ_1, ϕ_2) is decomposed as

$$(\phi_1, \phi_2) = (\phi_1', \phi_2') + (\rho_1, \rho_2)$$

where

$$(\rho_1, \rho_2) := \left(\rho, -\sqrt{r}\rho\right). \tag{13}$$

Proof. Define $\rho(x) := \phi_1(x) - \phi_1'(x)$. Because $(\phi_1' + \rho_1, \phi_2' + \rho_2)$ must satisfy Eq. (10), it must hold that

$$(\phi_1' + \rho_1) = J_{\tilde{\alpha}}^{-1} u_\alpha - \frac{1}{\sqrt{r}} (\phi_2' + \rho_2).$$

Because ϕ_1' is an IIF, we have

$$(J_{\tilde{\alpha}}^{-1} u_\alpha - \frac{1}{\sqrt{r}} \phi_2'(x) + \rho_1) = J_{\tilde{\alpha}}^{-1} u_\alpha - \frac{1}{\sqrt{r}} (\phi_2' + \rho_2).$$

Thus, we obtain $\rho_1 + \frac{1}{\sqrt{r}} \rho_2 = 0$. To guarantee the finiteness of (ρ_1, ρ_2)'s covariance, it suffices to show that for any $z \in \Re^d$ $z^T E[(\phi_1 - \phi_1')(\phi_1 - \phi_1')^T] z$ is finite. Note that

$$z^T E[(\phi_1 - \phi_1')(\phi_1 - \phi_1')^T] z = z^T E[\phi_1 \phi_1^T] z + z^T E[\phi_1 (\phi_1')^T] z - z^T E[\phi_1 (\phi_1')^T] z - z^T E[\phi_1' \phi_1^T] z.$$

Because $E[\phi_1 \phi_1^T]$ and $E[\phi_1' (\phi_1')^T]$ are finite, the first and second terms are finite. As for the remaining term, we can show the finiteness similarly by using Cauchy Schwartz inequality. □

From this lemma, it is immediate to see that such a pair (ρ_1, ρ_2) is neither IIF nor NIF because it does not satisfy neither Eq. (10) nor Eq. (12).

Assumption 1 (model enlargement). *Assume that $\mathscr{M}_x$ is correctly specified with $d' < d$. Then, we can enlarge the model up to d-dimension as*

$$\overline{\mathscr{M}}_x := \{\bar{g}(x;\eta) := g(x;\eta_1)\exp(\eta_2 T(x) - \psi(\eta)) \mid \eta_1 \in \Re^{d'},\ \eta_2 \in \Re^{d-d'},\ \eta = (\eta_1^T, \eta_2^T)^T\}$$

where $\psi(\eta) := \log(\int g(x;\eta_1)\exp(\eta_2^T T(x))dx)$. Clearly, this model also contains $p(x)$ and has a score function

$$s_\eta(x;\eta) = \left(\frac{\partial \log g(x;\eta_1)}{\partial \eta_1}^T - \int \bar{g}(x;\eta)\frac{\partial \log g(x;\eta)}{\partial \eta}dx,\ T(x) - \int \bar{g}(x;\eta)T(x)dx\right).$$

Lemma 9. *Assume that $d \le d'$. Let $\rho : \mathscr{X} \to \Re^d$ such that $E_p[\rho(X)] = \mathbf{0}$ and ρ has a finite covariance matrix. Then, for any given full-rank matrix $A \in \Re^{d\times d'}$, there is $\nu : \mathscr{X} \to \Re^{d'}$ such that $E_p[\nu(X)] = \mathbf{0}$ and ν has a proper covariance matrix and $\rho(x) = A\nu(x)$.*

Proof. When $d \le d'$, let us prepare an arbitrary function $\nu_2 : \mathscr{X} \to \Re^{d'-d}$ such that $E[\nu_2(X)] = 0$ and ν_2 has a proper covariance matrix and linearly independent of ρ. Without loss of generality, A is decomposed as $A = [A_1\ A_2]$ where $A_1 \in \Re^{d\times d}$ is non-singular and $A_2 \in \Re^{d\times(d'-d)}$. Define

$$\nu_1(x) := A_1^{-1}\rho(x) - A_1^{-1}A_2\nu_2(x).$$

Then, $\nu(x) := (\nu_1(x)^T, \nu_2(x)^T)^T$. It is easy to see that

$$A\nu(x) = [A_1\ A_2]\begin{pmatrix}\nu_1(x)\\ \nu_2(x)\end{pmatrix} = A_1\nu_1(x) + A_2\nu_2(x) = \rho(x) - A_2\nu_2(x) + A_2\nu_2(x) = \rho(x).$$

Clearly, $E_{p(x)}[\nu(x)] = \mathbf{0}$. By construction, $E[\rho\rho^T]$ is proper because $E[\rho\rho^T]$ and $E[\nu_2\nu_2^T]$ are finite and $\nu(x)$ is linearly independent of each other. □

Lemma 10. *Let $\nu(x) : \mathscr{X} \to \Re^d$ be a function of x such that $E_{p(x)}[\nu(X)] = 0$ and $\nu(x)$ has a proper covariance matrix. Then, for any η_0, there is $u_\eta(x;\eta)$ such that*

$$\nu(x) = -E_{p(x)}\left[\frac{\partial u_\eta(X;\eta_0)}{\partial \eta^T}\right]^{-1} u_\eta(x;\eta_0)$$

and

$$E_{p(x)}[u_\eta(X;\eta_0)] = \mathbf{0}.$$

Proof. Let $V := E_{p(x)}[\nu(X)\nu(X)^T]$. By assumption, V is non-singular. Define

$$\bar{\nu}(x) := V^{-1}\nu(x).$$

This $\bar{\nu}(x)$ is a dual basis of $\nu(x)$, i.e.,

$$E[\nu(X)\bar{\nu}(X)^T] = E[\nu(X)\nu(X)^T]V^{-T} = VV^{-1} = I.$$

Let C be any fixed non-singular $(d \times d)$ matrix. Let further $a : \mathscr{X} \to \Re^d$ be a vector-valued function such that each component is perpendicular to $\mathrm{Span}(\nu(x))$. Using these, define

$$C(x;\eta) := \exp\left((a(x) - \bar{\nu}(x))^T(\eta - \eta_0)\right)C.$$

We write the k-th row of C and $C(x;\eta)$ as c_k and $c_k(x;\eta)$, i.e.,

$$C = \begin{bmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_d^T \end{bmatrix}, \; C(x;\eta) = \begin{bmatrix} c_1(x;\eta)^T \\ c_2(x;\eta)^T \\ \vdots \\ c_d(x;\eta)^T \end{bmatrix}.$$

By the chain rule of differential, we have

$$\frac{\partial u_\eta(x;\eta)}{\partial \eta^T} = \begin{bmatrix} \nu(x)^T \frac{\partial c_1(x;\eta)}{\partial \eta^T} \\ \nu(x)^T \frac{\partial c_2(x;\eta)}{\partial \eta^T} \\ \vdots \\ \nu(x)^T \frac{\partial c_d(x;\eta)}{\partial \eta^T} \end{bmatrix}.$$

Using these, define

$$u_\eta(x;\eta) := C(x;\eta)\nu(x).$$

First, noting that $u_\eta(x;\eta_0) = C\nu(x)$. we have

$$E_{p(x)}[u_\eta(X;\eta_0)] = CE_{p(x)}[\nu(X)] = \mathbf{0}.$$

Next, we calculate $\partial u_\eta / \partial \eta^T$. From its definition,

$$\frac{\partial c_k(x;\eta)}{\partial \eta^T} = \exp((a(x) - \bar{\nu}(x))^T(\eta - \eta_0))c_k(a(x) - \bar{\nu}(x))^T.$$

Taking its expectation with respect to $p(x)$ at $\eta = \eta_0$, we can calculate the k-th row of $E[\partial u_\eta(X;\eta_0)/\partial \eta^T]$ as

$$\begin{aligned} &= E_{p(x)}[(\nu(X)^T c_k)(a(X) - \bar{\nu}(X))^T] = E_{p(x)}\left[(c_k^T \nu(X))(a(X) - \bar{\nu}(X))^T\right] \\ &= -c_k^T E_{p(x)}\left[\nu(X)\bar{\nu}(X)^T\right] = -c_k^T I_d = -c_k^T. \end{aligned}$$

Therefore, we have

$$E\left[\frac{\partial u_\eta(x;\eta_0)}{\partial \eta^T}\right] = -C.$$

As a result,

$$-E_{p(x)}\left[\frac{\partial u_\eta(X;\eta_0)}{\partial \eta^T}\right]^{-1} u_\eta(x;\eta_0) = -(-C)^{-1}C\nu(x) = \nu(x).$$

□

Lemma 11 (scale decomposition). *Let $\nu(x)$ be an influence function of $\hat{\eta}$,i.e.,*

$$\sqrt{n^*}(\hat{\eta} - \eta_0) = \frac{1}{\sqrt{n^*}} \sum_{i=1}^{n^*} \nu(x_i) + o_p(1)$$

for any data number n^*. Depending on which data set are used, its influence function for $\sqrt{n}(\hat{\eta} - \eta_0)$ varies as written as

only labeled data	$(1,0) \cdot \nu(x)$
only unlabeled data	$(0, \sqrt{r}) \cdot \nu(x)$
both labeled/unlabeled data	$\left(\frac{r}{1+r}, \frac{\sqrt{r}}{1+r}\right) \cdot \nu(x)$

where $(1,0)\cdot\nu(x)$ *denotes* $(\nu(x),0)$ *for example. Next, consider a pair of function* $a\cdot(1,-\sqrt{r})\cdot\nu(x)$ *where* a *is a some real number. We can decompose it into the above terms' difference for some constants:*

$$\begin{aligned}
1\cdot(1,-\sqrt{r}) &= (1,0)-(0,\sqrt{r})\\
-1\cdot(1,-\sqrt{r}) &= (0,\sqrt{r})-(1,0)\\
\frac{1}{r+1}\cdot(1,-\sqrt{r}) &= (1,0)-\left(\frac{r}{r+1},\frac{\sqrt{r}}{r+1}\right)\\
-\frac{1}{r+1}\cdot(1,-\sqrt{r}) &= \left(\frac{r}{r+1},\frac{\sqrt{r}}{r+1}\right)-(1,0)\\
\frac{r}{r+1}\cdot(1,-\sqrt{r}) &= \left(\frac{r}{r+1},\frac{\sqrt{r}}{r+1}\right)-(0,\sqrt{r})\\
-\frac{r}{r+1}\cdot(1,-\sqrt{r}) &= (0,\sqrt{r})-\left(\frac{r}{r+1},\frac{\sqrt{r}}{r+1}\right)
\end{aligned}$$

where $\cdot\nu(x)$ *is omitted in each term for simplicity.*

Proof. When only the labeled data are used, it is trivial. When only the unlabeled data are used, we have

$$\sqrt{n'}(\hat{\eta}-\eta_0)=\frac{1}{\sqrt{n'}}\sum_{j=1}^{n'}\nu(x'_j)+o_p(1).$$

Because $\sqrt{r}\sqrt{n'}=\sqrt{n}$,

$$\sqrt{n}(\hat{\eta}-\eta_0)=\frac{1}{\sqrt{n'}}\sum_{j=1}^{n'}\frac{1}{\sqrt{r}}\nu(x'_j)+o_p(1).$$

Similarly, when both the labeled data and unlabeled data are used, we have

$$\sqrt{n+n'}(\hat{\eta}-\eta_0)=\frac{1}{\sqrt{n+n'}}\left(\sum_{i=1}^{n}\nu(x_i)+\sum_{j=1}^{n'}\nu(x'_j)\right)+o_p(1).$$

By multiplying $\sqrt{n}/\sqrt{n+n'}$,

$$\sqrt{n}(\hat{\eta}-\eta_0)=\frac{\sqrt{n}}{n+n'}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{n}\nu(x_i)+\frac{1}{\sqrt{n'}}\sum_{j=1}^{n'}\sqrt{n'}\nu(x'_j)\right)+o_p(1).$$

It is immediate to confirm the decomposition from the above results. □

We say that an influence function of $\hat{\alpha}$ is a SIF if its influence function has the form $(J_{\tilde{\alpha}}^{-1}u_\alpha(x,y;\alpha_0),0)$. Clearly, this corresponds to a some supervised estimator.

Theorem 12 (decomposition theorem). *Assume that* $d\leq d'$. *Suppose that we are given an interesting influence function* $(\phi_1(x,y),\phi_2(x))$. *For any IIF* (ϕ'_1,ϕ'_2), *any* $(d\times d')$ *full-rank matrix* A *and nuisance parameter value* η_0, *there exist two nuisance influence functions associated with* η_0 *such that* (ϕ_1,ϕ_2) *is decomposed as*

$$(\phi_1,\phi_2)=(\phi'_1,\phi'_2)+A(\nu_1,\nu_2)-A(\nu'_1,\nu'_2). \tag{14}$$

and two NIFs satisfy Eq. (15) in the proof.

Proof. By the series of previous lemmas, it is almost trivial. By Lemma 8 with an arbitrary IIF $((\phi_1', \phi_2'))$, there exists function $\rho(x)$ such that $E_p[\rho(X)] = 0$ and $E_p[\rho(X)\rho(X)^T] < \infty$ and

$$(\phi_1, \phi_2) = (\phi_1', \phi_2') + \left(\rho(x), -\sqrt{r}\rho(x)\right).$$

By Lemma 9, for any full-rank $(d \times d')$ matrix A, there exists $\nu(x)$ such that $E_p(x)[\nu(X)] = \mathbf{0}$ and ν has a proper covariance matrix. Hence, we have

$$(\phi_1, \phi_2) = (\phi_1', \phi_2') + A\left(\nu(x), -\sqrt{r}\nu(x)\right).$$

Otherwise, by Lemma 10, there exists $u_\eta(x;\eta) = C(x;\eta)\nu(x)$ with a given η_0 such that

$$-E\left[\frac{\partial u_\eta(X;\eta_0)}{\partial \eta^T}\right] u_\eta(X;\eta_0) = \nu(x)$$

and

$$E[u_\eta(X;\eta_0)] = \mathbf{0}.$$

For any NIF $(\nu_1'(x), \nu_2'(x))$ associated with η_0 and $u_\eta(x;\eta)$, define

$$(\tilde{\nu}_1(x), \tilde{\nu}_2(x)) := (\nu(x), -\sqrt{r}\nu(x)) + (\nu_1'(x), \nu_2'(x)).$$

Then, they satisfy

$$(\nu, -\sqrt{r}\nu) = (\tilde{\nu}_1, \tilde{\nu}_2) - (\nu_1', \nu_2'). \tag{15}$$

This pair of function $(\tilde{\nu}_1, \tilde{\nu}_2)$ satisfies

$$\tilde{\nu}_1(x) + \frac{1}{\sqrt{r}}\tilde{\nu}_2(x) = \nu(x) + \nu_1'(x) + \frac{1}{\sqrt{r}}\left(-\sqrt{r}\nu(x) + \nu_2'(x)\right) = \nu_1'(x) + \frac{1}{\sqrt{r}}\nu_2'(x) = J_{\tilde{\eta}}^{-1} u_\eta(x;\eta_0)$$

by Theorem 7. This indicates that $(\tilde{\nu}_1(x), \tilde{\nu}_2(x))$ is also another NIF. □

4 Construction of Estimating Equations

By the decomposition theorem, we know that any IIF consists of an IIF and two NIFs. Using this fact, let us specify estimating equations yielding the estimator with the given IIF. We immediately have the following corollary from the decomposition theorem:

Theorem 13. *In semi-supervised setting, suppose that we are given an interest function (ϕ_1, ϕ_2) of regular and asymptotically linear estimator $\hat{\alpha}$. If the estimating function corresponding to the supervised influence function is available (written as μ_α), there exist two estimating functions μ_η and $\mu_{\eta'}$ such that the solution of the following equation*

$$\begin{aligned}
\sum_{i=1}^{n} \mu_\alpha(x_i, y_i; \alpha) f(w(x_i; \eta, \eta')) &= \mathbf{0}, \\
\sum_{i=1}^{n} \mu_\eta(x_i, y_i; \eta) + \sum_{j=1}^{n'} \mu_\eta(x_j'; \eta) &= \mathbf{0}, \\
\sum_{i=1}^{n} \mu_\eta'(x_i, y_i; \eta') + \sum_{j=1}^{n'} \mu_\eta'(x_j'; \eta') &= \mathbf{0}.
\end{aligned} \tag{16}$$

has the given interest influence function. Here, f is any function such that $f(1) = 1$ and $f'(1) = 1$.

Proof. Note that

$$\begin{aligned}\frac{\partial f(w(x;\eta,\eta'))}{\partial \eta} &= -f'(w(x;\eta,\eta'))w(x;\eta,\eta')s_\eta(x;\eta),\\ \frac{\partial f(w(x;\eta,\eta'))}{\partial \eta'} &= f'(w(x;\eta,\eta'))w(x;\eta,\eta')s_\eta(x;\eta'),\end{aligned}$$

By Taylor-expansion around (α_0,η_0,η_0), we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial \mu_\alpha(x_i,y_i;\alpha_0)}{\partial \alpha^T}\sqrt{n}(\hat{\alpha}-\alpha_0)\\ &\quad + \frac{1}{n}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0)\left(\frac{\partial f(w(x_i;\eta_0,\eta_0))}{\partial \eta}\right)^T \sqrt{n}(\hat{\eta}-\eta_0)\\ &\quad + \frac{1}{n}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0)\left(\frac{\partial f(w(x_i;\eta_0,\eta_0))}{\partial \eta'}\right)^T \sqrt{n}(\hat{\eta}'-\eta_0)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial \mu_\alpha(x_i,y_i;\alpha_0)}{\partial \alpha^T}\sqrt{n}(\hat{\alpha}-\alpha_0)\\ &\quad + \frac{1}{n}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0)s_\eta(x_i;\eta_0)^T\sqrt{n}(\hat{\eta}'-\eta_0) - \frac{1}{n}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0)s_\eta(x_i;\eta_0)^T\sqrt{n}(\hat{\eta}-\eta_0)\\ &\xrightarrow{P} \frac{1}{\sqrt{n}}\sum_{i=1}^n \mu_\alpha(x_i,y_i;\alpha_0) - J_{\tilde{\alpha}}\sqrt{n}(\hat{\alpha}-\alpha_0) + G_{\tilde{\alpha}\eta}\sqrt{n}(\hat{\eta}'-\eta_0) - G_{\tilde{\alpha}\eta}\sqrt{n}(\hat{\eta}-\eta_0).\end{aligned}$$

Therefore, we have

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n J_{\tilde{\alpha}}^{-1}\mu_\alpha(x_i,y_i;\alpha_0) + J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}\sqrt{n}(\hat{\eta}'-\eta_0) - J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}\sqrt{n}(\hat{\eta}-\eta_0).$$

By Theorem 12 with $A = J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}$, $\eta_0 = 0$ and $(\phi_1',\phi_2') = J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta}$, (ϕ_1,ϕ_2) can be decomposed as

$$(\phi_1,\phi_2) = (J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\eta},0) + A(\nu_1,\nu_2) - A(\nu_1',\nu_2')$$

where (ν_1,ν_2) and (ν_1',ν_2') are influence functions associated with estimating functions $u_\eta(x;\eta) = \exp(-\eta^T\bar{\nu}(x))\nu(x)$ and $u_\eta'(x;\eta) = \sqrt{r}\exp(-\eta'^T\bar{\nu}(x))\nu(x)$. This form is equal to the IIF given by solving Eq. (16). □

Using this theorem, we can recover the estimating equation of DRESS I and DRESS II from their influence functions. However, their recovery is not unique. We cannot guarantee the recovery of the original estimating equations without any prior knowledge about them.

5 Effective and Efficient Class of Influence Function and Its Associated Estimators

We derive the most efficient influence functions associated with regular and asymptotic linear estimators.

Theorem 14. *Let (ϕ_1, ϕ_2) be an arbitrary interest influence function of RAL estimator. The IIF (ϕ_1, ϕ_2) improves (overcomes or is equal to) the supervised influence function $J_{\tilde{\alpha}}^{-1} u_\alpha(x, y; \alpha)$ with respect to the asymptotic variance of the associated estimator if and only if*

$$\begin{aligned} 0 \leq & \frac{1}{\sqrt{r}} \left(J_{\tilde{\alpha}}^{-1} E_{p(x)}[\bar{u}_\alpha(X;\alpha_0)\phi_2(X)^T] + E_{p(x)}[\phi_2(X)\bar{u}_\alpha(X;\alpha_0)^T] J_{\tilde{\alpha}}^{-T} \right) \\ & - \left(1 + \frac{1}{r}\right) E_{p(x)}[\phi_2(X)\phi_2(X)]. \end{aligned}$$

where $\bar{u}_\alpha(x;\alpha) := E_{p(y|x)}[u_\alpha(X, Y; \alpha)|X = x]$.

Proof. For any IIF (ϕ_1, ϕ_2), the asymptotic covariance of its associated estimator is calculated as

$$\begin{aligned} \mathrm{Avar}(\tilde{\alpha}) &= E_{p(x,y)}[\phi_1(X,Y)\phi_1(X,Y)^T] + E_{p(x')}[\phi_2(X')\phi_2(X')^T] \\ &= E\left[\left(J_{\tilde{\alpha}}^{-1} u_\alpha(X,Y;\alpha_0) - \frac{1}{\sqrt{r}}\phi_2(X) \right) \left(J_{\tilde{\alpha}}^{-1} u_\alpha(X,Y;\alpha_0) - \frac{1}{\sqrt{r}}\phi_2(X) \right)^T \right] \\ &\quad + E_{p(x')}[\phi_2(X')\phi_2(X')^T] \\ &= J_{\tilde{\alpha}}^{-1} G_{\tilde{\alpha}\tilde{\alpha}} J_{\tilde{\alpha}}^{-T} - \frac{1}{\sqrt{r}} E\left[\phi_2(X) u_\alpha(X,Y;\alpha_0)^T J_{\tilde{\alpha}}^{-T}\right] - \frac{1}{\sqrt{r}} E\left[J_{\tilde{\alpha}}^{-1} u_\alpha(X,Y;\alpha_0)\phi_2(X)^T\right] \\ &\quad + \frac{1}{r} E_{p(x)}\left[\phi_2(X)\phi_2(X)\right] + E_{p(x')}[\phi_2(X')\phi_2(X')^T] \end{aligned}$$

where $\mathrm{Avar}(\hat{\alpha})$ is defined as the variance of the limit distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$. Because the first term is just the asymptotic covariance of supervised estimator, we obtain the statement. □

Theorem 15. *The most efficient interest influence function is*

$$(\phi_1, \phi_2) = \left(J_\alpha^{-1} u_\alpha - \frac{1}{r+1} J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(x;\alpha_0), \frac{\sqrt{r}}{r+1} J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(x';\alpha_0) \right).$$

Proof. Let $z \in \Re^d$ be any real vector. Define

$$\begin{aligned} H(\phi_2) &:= z^T \Big(J_{\tilde{\alpha}}^{-1} E[\bar{u}_\alpha(X;\alpha_0)\phi_2(X)^T] + E[\phi_2(X)\bar{u}_\alpha(X;\alpha_0)^T] J_{\tilde{\alpha}}^{-T} \\ &\quad - \frac{r+1}{\sqrt{r}} E[\phi_2(X)\phi_2(X)] \Big) z \\ &= E\left[z^T J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(X;\alpha_0)\phi_2(X)^T z + z^T \phi_2(X)\bar{u}_\alpha(X;\alpha_0)^T J_{\tilde{\alpha}}^{-T} z - \frac{r+1}{\sqrt{r}} z^T \phi_2(X)\phi_2(X) z \right] \\ &= E\left[2 z^T J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(X;\alpha_0)\phi_2(X)^T z - \frac{r+1}{\sqrt{r}} (z^T \phi_2(X))^2 \right]. \end{aligned}$$

By variation method with respect to $\phi_2(x)$, we have

$$\left(2 z^T J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(X;\alpha_0) - 2\frac{r+1}{\sqrt{r}} (z^T \phi_2(X)) \right) z = 0$$

for any z. Therefore, ϕ_2 is

$$\frac{\sqrt{r}}{r+1} J_{\tilde{\alpha}}^{-1} \bar{u}_\alpha(X;\alpha_0).$$

□

In this case, the asymptotic covariance is calculated as follows. Let

$$\bar{G}_{\tilde{\alpha}} := E_{p(x)}[\bar{u}_\alpha(X;\alpha_0)\bar{u}_\alpha(X;\alpha_0)^T].$$

It is straightforward to see that

$$\begin{aligned}
E[u\bar{u}_\alpha^T] &= E[\bar{u}_\alpha\bar{u}_\alpha^T] = \bar{G}_{\tilde{\alpha}} \\
E\left[J_{\tilde{\alpha}}^{-1}u_\alpha(X,Y;\alpha_0)\left(-\frac{1}{\sqrt{r}}\phi_2(X)^T\right)\right] &= -\frac{1}{r+1}E\left[J_{\tilde{\alpha}}^{-1}u_\alpha\bar{u}_\alpha^T J_{\tilde{\alpha}}^{-T}\right] = -\frac{1}{r+1}J_{\tilde{\alpha}}^{-1}\bar{G}_{\tilde{\alpha}}J_{\tilde{\alpha}}^{-T} \\
E\left[\left(-\frac{1}{\sqrt{r}}\phi_2(X)\right)\left(-\frac{1}{\sqrt{r}}\phi_2(X)\right)^T\right] &= \frac{1}{(r+1)^2}E\left[J_{\tilde{\alpha}}^{-1}\bar{u}_\alpha\bar{u}_\alpha^T J_{\tilde{\alpha}}^{-T}\right] = \frac{1}{(r+1)^2}J_{\tilde{\alpha}}^{-1}\bar{G}_{\tilde{\alpha}}J_{\tilde{\alpha}}^{-T}.
\end{aligned}$$

Using these, we have

$$\begin{aligned}
E[\phi_1\phi_1^T] + E[\phi_2\phi_2^T] &= J_{\tilde{\alpha}}^{-1}G_{\tilde{\alpha}\tilde{\alpha}}J_{\tilde{\alpha}}^{-T} + \left(-\frac{2}{r+1} + \frac{1}{(r+1)^2} + \left(\frac{\sqrt{r}}{r+1}\right)^2\right)J_{\tilde{\alpha}}^{-1}\bar{G}_{\tilde{\alpha}\tilde{\alpha}}J_{\tilde{\alpha}}^{-T} \\
&= J_{\tilde{\alpha}}^{-1}\left(G_{\tilde{\alpha}\tilde{\alpha}} - \frac{1}{r+1}\bar{G}_{\tilde{\alpha}\tilde{\alpha}}\right)J_{\tilde{\alpha}}^{-T}.
\end{aligned}$$

6 Conclusion

We specified the set of all possible influence functions of regular and asymptotic linear semi-supervised estimator under the conditional misspecification. We also showed that DRESS type estimating equation is universal because, for any given influence function of regular and asymptotic linear estimator, this type of estimating equation can yield an estimator having the given influence function. However, this construction is less useful. In real situations, we are not given a fixed influence function but given an influence function as a functional of joint distribution. It is valuable to extend our result to this case.

References

Kawakita, M., Takeuchi, J., May 2014. Safe semi-supervised learning based on weighted likelihood. Neural Networks 53, 146–164.

Tsiatis, A., 2006. Semiparametric Theory and Missing Data. Springer.